AFHRL-TP-89-33

2

AIR FORCE

HUMAN RESOURCES

LABORATORY

AIR FORCE OFFICER QUALIFYING TEST (AFOQT): DEVELOPMENT OF AN ITEM BANK

Willa B. Gupta
Frances R. Berger
Raymond M. Berger

Psychometrics, Incorporated
13245 Riverside Drive
Sherman Oaks, California 91423-2172

Jacobina Skinner

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

December 1989
Interim Technical Paper for Period January 1988 - July 1989

Approved for public release; distribution is unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

This paper has been reviewed and is approved for publication.

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

DANIEL L. LEIGHTON, Colonel, USAF
Chief, Manpower and Personnel Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
December 1989

3. REPORT TYPE AND DATES COVERED
Interim -- January 1988 to July 1989

4. TITLE AND SUBTITLE
Air Force Officer Qualifying Test (AFOQT):
Development of an Item Bank

5. FUNDING NUMBERS

C	- F33615-83-C-0035
PE	- 62703F
PR	- 7719
TA	- 18
WU	- 24

6. AUTHOR(S)

Willa B. Gupta   Raymond M. Berger
Frances R. Berger   Jacobina Skinner

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Psychometrics, Incorporated
13245 Riverside Drive
Sherman Oaks, California 91423-2172

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFHRL-TP-89-33

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The Air Force Officer Qualifying Test (AFOQT) has been part of the selection process for officer commissioning programs and pilot and navigator training since 1951. Form O, the latest of 15 successive forms of the AFOQT, contains 380 items organized into 16 subtests which form five composites: Pilot, Navigator-Technical, Academic Aptitude, Verbal, and Quantitative. The anticipated need for future forms of the AFOQT prompted the development of a large pool of experimental items. Approximately 6,000 items were developed and administered (along with Form O items) to basic airmen and officer cadets attending military training programs. One of the goals of this development was to provide an item bank for the development of future forms that emulate the content of Form O. The enhanced identification system accompanying each item allows increased efficiency of data retrieval to locate and combine desired item sets. This was accomplished by the creation of a magnetic data tape containing statistics, keys, sample identification, and information for interfacing tape data with item text residing on a computer data bank, with item graphics on cards, and with the printed experimental test booklets. The new AFOQT item bank contains the essential components for automated test construction. The next step in providing a sophisticated item bank would be to employ a computer application that permits immediate terminal display of item text, illustrations, graphs and statistics.

14. SUBJECT TERMS

Air Force Officer Qualifying Test	item banking
aptitude tests	mental abilities testing
CATC	officer selection (Continued)

15. NUMBER OF PAGES
42

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
UL

Item 14 (Concluded):

officer classification
test construction

AFHRL Technical Paper 89-33 December 1989

AIR FORCE OFFICER QUALIFYING TEST (AFOQT): DEVELOPMENT OF AN ITEM BANK

Willa B. Gupta
Frances R. Berger
Raymond M. Berger

Psychometrics Inc.
13245 Riverside Drive
Sherman Oaks, California 91423-2172

Jacobina Skinner

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed by

Lonnie D. Valentine, Jr.
Chief, Force Acquisition Branch

Submitted for publication by

Lonnie D. Valentine, Jr.
Chief, Force Acquisition Branch

This publication is primarily a working paper. It is published solely to document work performed.

SUMMARY

This paper presents the steps in the development of an item bank to be used in computer-assisted test construction (CATC) of future Air Force Officer Qualifying Test (AFOQT) forms. The selection process for officer commissioning programs and pilot and navigator training has included the AFOQT since 1951. The importance of this use requires periodic checks on the AFOQT's predictive validity, currency, and security. These in turn may determine the need for new forms of the battery. Anticipating this need, the Air Force Human Resources Laboratory (AFHRL) initiated a project to develop a large pool of new items from which to draw new forms of the AFOQT and to design an efficient item banking and retrieval procedure.

Historically, the texts, statistics, and sample information associated with individual experimental items were recorded on cards. These were referenced manually when considering experimental items for use in operational tests. The upgrading of this cumbersome and time-consuming system was considered necessary to increase the efficiency of locating and combining items with desirable features, as well as to enhance data security. The item bank was designed to satisfy these needs and to ensure compatibility with data retrieval systems associated with CATC.

The item storage systems developed for this project link item text, item data, and item graphics to facilitate the locating and combining of items for future AFOQT forms. A set of floppy diskettes contains the complete text of non-pictorial items and a card deck contains the graphic items. A data tape contains various item identification codes, sample identification codes, and statistical data. The test statistics include the classical item analysis data with biserial correlations and item difficulty, quintiles, and logistic item response theory analysis data. The information for identifying item records on the tape and linking these to item texts and the illustrations in the card deck is described.

The next step in providing sophisticated access to the AFOQT item bank would be to employ a computer application that would enable all text, statistics, graphs and illustrations to be viewed on the terminal screen using a few simple commands.

The appendixes to this paper include a taxonomy of the AFOQT content areas, rules for item writing, the tape file layout, and steps in the construction of the data bank tape.

PREFACE

The Air Force Human Resources Laboratory (AFHRL) is designated as the primary test development agency for the Air Force Officer Qualifying Test (AFOQT) by Air Force Regulation 35-8, Air Force Military Personnel Testing System. The current research and development (R&D) effort was undertaken as part of AFHRL's responsibility to develop, revise, and conduct research in support of the AFOQT. Work was accomplished under Task 771918, Selection and Classification Technologies, which is part of a larger effort in Force Acquisition and Distribution Systems. The project was completed under work unit 77191824, Officer Item Pool Development, as a contractual effort (Contract F33615-83-C-0035) by Psychometrics Inc., Sherman Oaks, CA. Ms. Jacobina Skinner represented the monitoring agency as contract manager.

The authors wish to express their appreciation of the scientific and technical guidance given by Drs. William E. Alley, Lonnie D. Valentine, Jr., and Malcolm James Ree. We wish also to express our thanks to current and former AFHRL behavioral scientists including Dr. Janos B. Koplyay, Major John Quebe, 1Lt Thomas O. Arth, and Mr. Douglas Cowan for their informative review of items and test booklets.

The test administration support by Mr. Roy Chollman, Mr. Refugio Gonzales, and other test examiners at the AFHRL test facility at Lackland AFB, Texas is greatly appreciated, as is the item and test analysis support by Ms. Doris Black, Mr. Bill Glasscock, and other staff members of the AFHRL Information Sciences Division.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF CHARTS

Chart Page

AIR FORCE OFFICER QUALIFYING TEST (AFOQT): DEVELOPMENT OF AN ITEM BANK

I. INTRODUCTION

One of the decision points in the officer candidate selection system for the United States Air Force is "mental qualification," as determined by scores on the Air Force Officer Qualifying Test (AFOQT). The AFOQT is used to select individuals for Officer Training School, Air Force Reserve Officer Training Corps cadets for scholarships or for the Professional Officers Course, and students for Undergraduate Pilot Training and Undergraduate Navigator Training. (Applicants to the Air Force Academy are exempt from this testing requirement.) Due to the importance of its use, the AFOQT requires periodic checks on its predictive validity, its currency, and its security, which in turn may determine the need for new forms of the battery. Anticipating this need, the Air Force Human Resources Laboratory (AFHRL) initiated a project to develop a large pool of new items from which new forms of the AFOQT could be assembled, and to design and implement a procedure for efficient banking and retrieval of these items.

The development of the pool of AFOQT items was aimed at emulating the style and content of Form O, the AFOQT in use for testing officer applicants at the start of this project. Form O is a multiple aptitude battery of 16 subtests covering verbal, quantitative, perceptual, and specialized ability areas. The subtest titles, number of items, and a brief description of the aptitudes and abilities measured are presented in Table 1. Readers interested in a more detailed description of test content and samples of the items are referred to the test manual for the AFOQT (Berger, Gupta, Berger, & Skinner, in preparation). The amount of time required to administer the entire battery is about 4.5 hours. Scores for the subtests are computed by summing the correct responses. Subtest raw scores are combined into five composite scores (Verbal, Quantitative, Academic Aptitude, Pilot, and Navigator-Technical) which are converted to percentiles for use in selecting and classifying officer applicants into military training programs.

The rationale for providing the continuity of content between Form O and the item pool arises from the value of Form O for predicting success in training. Validity studies have demonstrated that the composites correlate significantly with performance in the two officer commissioning programs for which the test is used as a selection factor: Officer Training School (Cowan, Barrett, & Wegner, 1990) and Reserve Officer Training Corps (Cowan, Barrett, & Wegner, 1989). Further, the subtests and composites are predictive of performance in follow-on specialized training courses for aircrew jobs (Arth, Steuck, Sorrentino, & Burke, in preparation) and for non-aircrew jobs (Arth, 1985; Arth & Skinner, 1986; Finegold & Rogers, 1985). A summary of the validity studies for Form O is also given in the AFOQT test manual (Berger et al., in preparation).

The pool of available items for officer test development was exhausted during the construction of Form O. A major purpose of the current project was to replenish the item pool to support the development of replacement tests for Form O. This was to include its immediate successors, Forms P1 and P2 (Berger, Gupta, Berger, & Skinner, 1988), as well as the next two or three generations of the test. Historically, the text, statistics, and other information needed on individual experimental items being considered for use in operational tests were recorded on cards. Test construction involved manually manipulating the card file, a time-consuming and cumbersome process.

The upgrading of the system of item storage at AFHRL was considered necessary for increasing the efficiency of data retrieval to locate and combine desired sets of items; for enhancing data security; and for ensuring compatibility with data retrieval systems involved with computerized test construction. As indicated by Muiznieks and Dennis (1979), the construction of parallel tests with given specifications and distribution characteristics is facilitated by automated item banking, and test security becomes less of a matter of concern when there are parallel items in the bank. Ree (1978) established that the use of an item banking system protects items against loss or compromise better than does storage on cards. Moreover, a test whose items are

Table 1. **Description of Items in AFOQT Form O Subtests**

Subtest	No. of items	Measures of aptitude/ability/knowledge
Verbal Analogies	25	Ability to reason and recognize relationships between words.
Arithmetic Reasoning	25	Ability to understand and reason with arithmetic relationships.
Reading Comprehension	25	Ability to read and understand paragraphs.
Data Interpretation	25	Ability to interpret data from graphs and charts.
Word Knowledge	25	Ability to understand written language through use of synonyms.
Math Knowledge	25	Ability to use learned mathematical terms, formulas, and relationships.
Mechanical Comprehension	20	Mechanical knowledge and understanding of mechanical functions.
Electrical Maze	20	Spatial ability to choose a correct path through a maze.
Scale Reading	40	Ability to read scales and dials.
Instrument Comprehension	20	Ability to determine aircraft attitude from flight instruments.
Block Counting	20	Spatial ability to "see into" a three-dimensional pile of blocks.
Table Reading	40	Ability to read tables quickly and accurately.
Aviation Information	20	Knowledge of general aeronautical concepts and terminology.
Rotated Blocks	15	Spatial aptitude by visualizing and manipulating objects in space.
General Science	20	Knowledge and understanding of scientific terms, concepts, principles, and instruments.
Hidden Figures	15	Perceptual and visual imagery ability using simple figures embedded in complex drawings.

computer selected and printed should result in greater uniformity of appearance and fewer typographical errors than one whose items are manually compiled. Large amounts of item data can be more quickly, reliably, and accurately stored, retrieved, and analyzed by use of an item bank tape than by use of card files.

The item bank developed for this project contains classical and Item Response Theory (IRT) statistics, sample information, and test form identifiers. Each item developed for an AFOQT subtest can be identified and linked to its statistical data and sample characteristics.

The next part of this paper summarizes the development of the new experimental item pool. Part III reviews the item banking procedures, and Part IV discusses issues concerning AFOQT item development and storage.

II. ITEM DEVELOPMENT

Taxonomy

Content and Style. One of the goals for item development was to be consistent with Form O in content and style. To create a foundation for distributing the items in terms of topics and stylistic features, the texts of Form O items, including stem and alternatives, and their item data were categorized for salient characteristics. Schemes for taxonomic classification were developed for items in 10 subtests. The nature of the items in the other six AFOQT subtests (Data Interpretation, Electrical Maze, Block Counting, Table Reading, Rotated Blocks, and Hidden Figures) did not lend to categorization. For the 10 subtests with taxonomies, the categories identified are shown in Appendix A. When appropriate, the content categories covered in selected subtests were expanded to include a broader concept of the area tested; for example, adding computer-related questions to the General Science subtest because of their importance in the science fields. The relative contribution of each category to a Form O subtest became the model guiding the new item construction.

Difficulty. Form O items were further analyzed to gain insight about factors that seemed to affect difficulty. These findings were then referenced in constructing items to approximate the same range of difficulty as Form O within each subtest. The sample on which item data were based consisted of all first-time Form O examinees tested for commissioning qualification between 1 March 1982 and 29 February 1984 (N = 75,980).

Rules for Item Writing

The development of the pool of new AFOQT items took into account the standard considerations for item writing, as delineated, for exampie, by Wesman (1971). Item writers, whether project staff or subject-matter consultants, were provided with descriptions of the requirements for content, scope, complexity, length, appearance, graphics, and number of items and response options. The principal consideration that guided the appearance and content of the new items was consistency with Form O. Emphasis was also placed on the need to be sensitive to issues of gender, ethnicity, regional characteristics, morals, politics, and religion. (If any of these topics was used, it was in a distanced and inoffensive way. For example, one of the Reading Comprehension items discussed archeological clues relating the migration of some English villages to the changing sites of churches.) Specific rules followed by item writers are shown in Appendix B.

Item Production

Phase 1. A total of 6,024 experimental items was written in two phases. In the initial phase, 301 new items were prepared for each subtest. As new items were constructed, they were pilot-tested informally by the contractor's staff and revised as necessary by the project directors. Further critiquing and editing were accomplished by AFHRL test specialists and the items were returned to the contractor for assembly into test booklets.

Phase 2. Additional experimental items were developed for selected subtests in a second phase. A project goal was to obtain, after field testing, approximately 175 new items per subtest meeting several statistical criteria. First, items needed to be within acceptable difficulty ranges. Second, items needed to differentiate among examinees of different ability levels. Item-total test (biserial) correlations for keyed responses higher than .40 were desired. Third, negative item-test correlations for incorrect responses were required. Statistical techniques used to evaluate item acceptability are described in the next section of this paper. Analysis results from the field test of items written in Phase 1 revealed which subtests would require item production in Phase 2.

Two types of items were prepared in the second phase. Some items were original, i.e., new in entirety, while others were revisions of items written in Phase 1. In some cases items were revised if they had acceptable item-test correlations but their difficulty levels were too easy or too hard. These items usually required shortening the stem or clarifying the meaning of the stem or alternatives. In other cases, the biserial correlations were acceptably high for the keyed response and negative for all but one wrong alternative. For these items, the wrong alternatives were made more clearly wrong if the biserial was positive and less transparently wrong if very few examinees were selecting that alternative. All items prepared in Phase 2 were subjected to the same editing, field testing, and statistical analysis as those prepared in Phase 1.

The total number of new items prepared for each subtest is shown in the first data column in Table 2. Item writing completed in Phase 1 resulted in a total of 301 new items for each of the sixteen subtests. The goal of 175 items meeting the acceptability criteria was achieved in eight of the subtests in Phase 1. Additional items were written in Phase 2 for the remaining eight of the sixteen subtests in numbers estimated to be required to meet the goal of 175 items meeting the statistical criteria of acceptability.

Experimental Test Booklets

From 7 to 16 booklets were assembled for each subtest, each booklet containing from 43 to 45 of the new test items (see Table 2). The number of test booklets required depended on the total number of experimental items written and on the power or speeded designation of the subtest. The 301 items prepared in Phase 1 were distributed among seven booklets for each subtest, except Scale Reading (SR) and Table Reading (TR) which had 14 booklets. SR and TR had been defined as speeded tests in Form O (Rogers, Roach, & Wegner, 1986). In order for item statistics for the last items of a speeded test to be as accurate as those for items that begin the test, it is important that all items be attempted. To ensure that this goal would be met, 14 booklets were prepared for these subtests, seven with items presented in forward order and seven with the same items presented in reverse order layouts. All items in power subtests were presented in forward order only. The booklets in excess of seven for all other subtests, which were power subtests, contained the additional new items prepared in Phase 2.

Common Items. In addition to the new items, each booklet contained a set of 15 to 20 items drawn from Form O (see Table 2). The items, which were the same for all booklets in a subtest, are referred to as common items. Common items were selected by AFHRL staff and provided a basis for verifying that the different samples tested on the various booklets of a subtest were comparable in terms of ability level. They were also useful for estimating the difficulty of new items for officer applicants. The common items followed the Form O order; that is, they were in the same location relative to each other in the experimental booklets as they were in the Form O test booklet. The selection of Form O common items was based on data from the operational sample of 75,980 applicants for officer commissioning training who took Form O between 1 March

Table 2. **Composition of Experimental Test Booklets**

Subtest	Total number of new items	Total number of booklets	Number of new items per booklet 1-7	8+	Number of common items per booklet
Verbal Analogies	526	12	43	45	20
Arithmetic Reasoning	301	7	43	--	20
Reading Comprehension	345	8	43	44	20
Data Interpretation	346	8	43	45	20
Word Knowledge	433	10	43	44	20
Math Knowledge	391	9	43	45	20
Mechanical Comprehension	697	16	43	44	19
Electrical Maze	301	7	43	--	20
Scale Reading	301	14	43	--	20
Instrument Comprehension	301	7	43	--	20
Block Counting	405	9	45	45	20
Table Reading	301	14	43	--	20
Aviation Information	301	7	43	--	20
Rotated Blocks	301	7	43	--	15
General Science	473	11	43	43	20
Hidden Figures	301	7	43	--	15
Total	6,024	153			

1982 and 29 February 1984. Items with optimum ranges of difficulty and discrimination were the first to be considered. Format, content, and graphic characteristics were also important considerations but were secondary to the statistical criteria.

Field Testing. Samples of approximately 350 basic airmen were tested on each booklet between August 1984 and December 1986 (Phase I) and between May 1986 and August 1988 (Phase 2) at Lackland Air Force Base, Texas. Several constraints precluded the use of the preferred sample of civilian applicants for Air Force officer commissions, the target population for which the AFOQT items were designed. Since the AFOQT is administered for operational selection and classification purposes at about 500 military testing sites in the Continental United States and overseas, it was not logistically or economically feasible to field test several thousand new test items with officer applicants. Basic airmen constituted the only practicable group on which to obtain preliminary data for evaluating item adequacy. The Basic Military Training (BMT) program has for many years provided a large and readily accessible source of examinees for AFHRL research and development (R&D) on skill and ability requirements for Air Force military occupations.

Supplemental data on items prepared in Phase 1 in two subtests - -- General Science (GS) and Aviation Information (AI) --- were obtained by readministering the seven experimental test booklets to samples of about 200 cadets attending Officer Training School (OTS) between October 1985 and January 1986. The GS and AI subtests assess knowledge in relatively technical and specialized areas. Results of the initial field tests suggested that airmen found the test content to be quite difficult. They answered from 29% to 36% of the items correctly in the various booklets. Item difficulty indices (proportion correct) fell below .30 for 41% to 65% of the items. These airmen performance levels prompted the establishment of special testing sessions with OTS cadets. Data obtained from the cadets were expected to provide a sounder basis for evaluating the adequacy of new items in the GS and AI subtests. OTS cadets are baccalaureate degree holders and have typically completed 2 to 4 more years of formal education than the majority of airmen.

Item data to augment those collected from basic airmen were obtained on selected booklets during Phase 2, also. The OTS cadet testing program continued with the administration of Mechanical Comprehension (booklets 8 - 11), another subtest containing items shown to be challenging for the basic airmen. In the summer of 1988, a third military personnel subgroup was tested. Cadets enrolled in the Reserve Officer Training Corps (ROTC) program were examined on the Mechanical Comprehension (booklets 12 - 16) and General Science (booklets 8 - 11) subtests. Testing occurred during field training encampments conducted at eight Air Force bases. Sample sizes ranged from about 110 to 250. The majority of ROTC cadets had completed their sophomore or junior year in college.

Test Administration. The collection of experimental item data for airmen was accomplished in test administration sessions lasting about 3 hours each. Multiple sessions were required to achieve the desired sample sizes for airmen. During each session it was typical for 45 airmen to be tested on 2 to 3 booklets. Each booklet contained items from a different subtest; this procedure ensured that the samples for different booklets in the same subtest were independent. Potential order-of-presentation effects were controlled by counterbalancing the sequence in which booklets were administered. Smaller groups of cadets were administered one content area booklet during a session.

Time limits for each power subtest were determined after the first few administrations of any test by noting the number of minutes required for 95% of the examinees to finish that subtest. The average became the time limit for the subsequent administration of the remaining booklets of that particular subtest. For the speeded subtests, the time limits were established based on the number of minutes required for 5% of the examinees to complete the tests.

The practices and procedures used to administer Form O at operational test sites were observed as closely as possible during collection of experimental item data. Major features of the manual for administration were replicated. For example, subtest directions were not changed. Demographics and test responses were recorded on a machine-scannable answer sheet (General Answer Sheet Type C, Westinghouse Corporation, Form 09 3937-001 W-2300).

III. THE ITEM BANK

One of the principal goals of the item construction project was to develop an item storage system that links three primary components --- item data, item text, and item graphics --- to facilitate the locating and combining of items for future AFOQT forms. To accomplish this goal, the essential data for each item were identified, a file layout was prepared (see Appendix C), and a tape containing the relevant information was designed and made ready for use. The variables described below are given on the item data tape. The procedural steps followed to create the tape are summarized in Appendix D. For narrative items, the complete text (stem and response options) was recorded separately on floppy diskettes. For items which reference graphics, a card deck was prepared containing a sample of the illustration.

Variables on the Data Tape

The descriptions of the coded variables are organized below according to content rather than to their sequential appearance on the data tape (see Appendix C).

Item Identification Codes

Content Area Identifier, a two-letter code, is an abbreviation of the subtest name, e.g., VA for Verbal Analogies.

Set Identifier, a direct reference to different booklets in each content area, starts with SET 1 for each content area.

Subject Type identifies the military personnel group used to field test the item (basic airmen, OTS cadets, or ROTC cadets).

Item Number refers to an item's location in a booklet. This variable has a range of 1 to 65.

Booklet Identifier is a unique five-digit number assigned to each subtest booklet.

Keyed Response, coded A to E for 15 subtests, and A to D for one subtest, stands for the correct answer.

Table 3 presents a summary of the coding structure for several item identification variables by AFOQT subtest.

Sample Identification Codes

Sample, a one-letter code, identifies the item as having been field tested by basic airmen (A), OTS cadets (C), or ROTC cadets (R). The Sample code is the same as the Subject Type part of the item identification information, but is in another position on the tape.

Sample Size, a 3-digit code, indicates the number of examinees field tested on the item.

Testing Date, an 8-digit code, indicates the month/year of the initial date of testing the item, and the month/year of the final date of testing.

Table 4 shows the range of codes for sample identification variables by AFOQT subtest.

Table 3. **Item Identification Codes**

Subtest	Content area	Sets	Booklets	Final item number
Verbal Analogies	VA	1 - 7	84001-84007	63
		8 - 12	85165-85169	65
Arithmetic Reasoning	AR	1 - 7	84008-84014	63
Reading Comprehension	RC	1 - 7	84015-84021	63
		8	85190	64
Data Interpretation	DI	1 - 7	84022-84028	63
		8	85189	65
Word Knowledge	WK	1 - 7	84029-84035	63
		8 - 10	85172-85174	64
Math Knowledge	MK	1 - 7	84036-84042	63
		8 - 9	85170-85171	65
Mechanical Comprehension	MC	1 - 7	84043-84049	62
		8 - 16	85195-85203	63
Electrical Maze	EM	1 - 7	84050-84056	63
Scale Reading	SR	1 - 7	84057-84063	63
		8 - 14	85153-85159	63
Instrument Comprehension	IC	1 - 7	84064-84070	63
Block Counting	BC	1 - 7	84071-84077	65
		8 - 9	85180-85181	65
Table Reading	TR	1 - 7	84078-84084	63
		8 - 14	84085-84091	63
Aviation Information	AI	1 - 7	84092-84098	63
Rotated Blocks	RB	1 - 7	84099-84105	58
General Science	GS	1 - 7	84106-84112	63
		8 - 11	85191-85194	63
Hidden Figures	HF	1 - 7	84113-84119	58

Table 4. **Sample Identification Codes**

Subtest	Content area	Sets	Sample	Range of sample sizes	Range of test dates
Verbal Analogies	VA	1 - 7	A	341-355	11-84 to 03-85
		8 - 12	A	350-400	11-86 to 08-87
Arithmetic Reasoning	AR	1 - 7	A	342-357	11-84 to 05-85
Reading Comprehension	RC	1 - 7	A	342-357	01-85 to 08-85
		8	A	348	08-87
Data Interpretation	DI	1 - 7	A	360-389	09-84 to 10-85
		8	A	345	08-87 to 08-88
Word Knowledge	WK	1 - 7	A	333-347	07-84 to 07-85
		8 - 10	A	400	11-86 to 03-87
Math Knowledge	MK	1 - 7	A	342-371	08-84 to 11-84
		8 - 9	A	400	11-86 to 03-87
Mechanical Comprehension	MC	1 - 7	A	351-367	09-84 to 12-85
		8 - 16	A	330-392	12-86 to 12-87
		8 - 11	C	188-224	09-87 to 08-88
		12 - 16	R	111-254	06-88 to 08-88
Electrical Maze	EM	1 - 7	A	340-375	09-84 to 09-85
Scale Reading	SR	1 - 7	A	342-357	06-85 to 08-85
		8 - 14	A	349-365	09-84 to 10-85
Instrument Comprehension	IC	1 - 7	A	337-583	11-84 to 02-85
Block Counting	BC	1 - 7	A	343-376	08-84 to 11-84
		8 - 9	A	349-373	07-87 to 08-88
Table Reading	TR	1 - 7	A	344-404	03-85 to 07-85
		8 - 14	A	344-398	01-85 to 07-85
Aviation Information	AI	1 - 7	A	336-356	12-84 to 04-85
		1 - 7	C	187-210	01-85 to 01-86
Rotated Blocks	RB	1 - 7	A	355-364	09-84 to 12-85
General Science	GS	1 - 7	A	348-359	01-85 to 07-85
		8 - 11	A	386-432	09-87 to 11-87
		1 - 7	C	198-207	12-84 to 01-86
		8 - 11	R	163-253	06-88 to 08-88
Hidden Figures	HF	1 - 7	A	333-349	07-84 to 08-84

Statistical Data on the Item Data Tape

Several analyses were conducted for each booklet set by sample type separately to evaluate the adequacy of the items in the experimental pool. Classical Item Statistics were derived using classical or "true score" theory analytic techinques (Gullikesen, 1950; Koplyay, 19¶1; Skinner & Ree, 1987). In addition, Officer Item Difficulty Estimates, IRT Item Statistics, and Quintile Statistics were obtained. Analysis results can be retrieved for any item by identifying the location of the variable from Appendix C, the File Layout.

Classical Item Analyses: Records 1 and 2. The biserial correlation (r_{bis}) between item score (correct or incorrect) and total test score (subtest raw score) was obtained as an index of item discrimination. The analyses made possible the identification of new items that reached or exceeded standards for distinguishing among examinees of differing ability levels. Primary requirements were that item-total score biserial correlations be positive and equal to or greater than .40 for keyed responses and be negative for all non-keyed (incorrect) alternatives. Only items meeting this standard were considered acceptable without revision for use in assembling future AFOQT forms. The number of items written for each subtest which met the item discrimination criterion is shown in Appendix E.

The item analysis also computed the percent of each sample responding correctly to each subtest item. This information was used to compare the ranges of difficulty of the new items with those of Form O and to point to where item reconstruction was necessary during Phase 2 of the item production part of the project.

As shown in Appendix C, Records 1 (R-1) and 2 (R-2) of the item bank tape contain the discrimination and difficulty statistics. The data are arranged for ease in reading on a terminal. Record 1 contains the observed item difficulty in columns 9 - 14. For each response option the biserial correlation (R-1) appears over the point-biserial correlation (R-2); percent choosing that response (R-1) appears over the number of examinees choosing that response (R-2); and the T value appears next (R-1). The T value is the mean score (standardized) on the total test of examinees selecting each item option.

Officer Item Difficulty Estimates. Early in the test construction project, concerns were raised as to how accurately item difficulty indices computed from responses of basic airmen and officer cadet subjects would reflect the actual difficulty on new items for officer applicants. The precision of item difficulty indices was questioned for two reasons. The first concerned the ability level of subjects and the second, an apparent speeded component underlying several subtests defined as power tests in AFOQT Form O (Rogers, Roach, & Wegner, 1986). The second issue is discussed in detail in the AFOQT Form P test construction technical report (Berger, Gupta, Berger, & Skinner, 1988).

It was anticipated that basic airmen would find the items more difficult on the average than would officer applicants, the majority of whom have completed more years of formal education. Conversely, the cadets, as a select group who had been previously screened and found to meet or exceed educational entry standards for the officer force as well as AFOQT score minimums, were expected to perform better than the larger pool of officer applicants. Multiple regression analyses were conducted to obtain weights needed to estimate how difficult new items tried out on basic airmen and cadets would be upon subsequent administration to officer applicants. Data for the criterion vector were difficulty indices for common items obtained from analyses of responses of 75,890 officer applicants administered AFOQT Form O under operational testing conditions between 1 March 1982 and 29 February 1984. Two types of information were included in the predictor set. Elements of the primary predictor vector for the corresponding common items were item difficulty values computed on airmen (or cadet) responses to experimental test booklets. A second predictor variable was developed to account for the potential relationship between the difficulty of an item and its location within the subtest. The location or position of each common item was recorded as its subtest item number in AFOQT Form O.

Analyses were conducted separately for each of 14 subtests in AFOQT Form O that had been treated specifically as power tests during the development and standardization of AFOQT Form O (Rogers, Roach, & Wegner, 1986). In the General Science and Aviation Information subtests, regression analyses were repeated on data collected from officer cadets to supplement the difficulty estimates obtained from basic airmen samples. Two equations were solved for each subtest and sample combination. One model constrained the relationship between item difficulty for officer applicants and basic airmen (or cadets) by item location to a linear form. Specifications for the second model permitted the relationship to take the more complex form of a curvilinear function. The total number of elements (N) for each model was equal to the number of common items in each subtest times seven, the number of independent airmen (and cadet) samples for which common difficulty values were available.

Inspection of squared multiple correlation coefficients (R^2) and standard errors of estimate (SEE) for the two models revealed that the data were adequately described by the more simple linear function. Follow-on analyses were conducted using the derived raw regression weights to compute estimates of the difficulty of new items (see Table 5).

Estimates of Item Difficulty: Record 3. This analysis contains values computed for each item for first, middle, and last item positions. The identified item positions are listed in Table 5.

IRT Analysis: Record 0. The a, b, and c parameter values derived from the IRT Analyses appear in Record 0. These statistics were computed with the Bilog II program (Mislevy & Bock, 1984).

Quintile Analysis Categorizing Information: Record 3. Quintile analysis requires distributions of subtest scores into quintiles. The total number of examinees were divided into five mutually exclusive score groups as equally as the data permitted. The quintile score boundaries varied from one booklet to another of the same subtest because of variations in score distributions among the different samples. A count was made of the number of examinees within a quintile selecting each response (including a "blank" category for omitted, unreadable, or double answers). Then, the percent of total examinees was calculated for each response within a quintile and across all quintiles.

The quintile data were used in the item evaluation and revision stage of the item production process. An advantage of using quintile information as a revision tool is that the data examined are raw data, not adjusted or normalized. Inferences were made concerning the quality of a response option by comparing obtained results to the ideal. Ideally, the lowest quintile will show equal numbers selecting the five alternatives, because the lowest scorers are assumed to be guessing. As the quintiles increase toward the top scores, an increasing number of people will select the keyed alternative. The breakpoint of random guesses and correct answers, which can occur between any two quintiles, adds to the information provided by the item difficulty index. A comparison of obtained data with the ideal will reveal how well the non-keyed options are performing as distractors or if there is ambiguity in the keyed option. If, for example, the same number of examinees select the same wrong answer in the first (lowest), second, and third quintiles, or if a high percentage of responses is found for the incorrect alternative in the fifth (highest) quintile, the need for revision is evident.

The lowest and highest score of each quintile and the corresponding number of examinees appear in Record 3.

Quintile Data: Records 4 through 9. Three variables are recorded for each response A to E and "no response": number of examinees in that quintile selecting that option; percent of examinees in that quintile selecting that option; and percent of total examinees selecting that option.

Table 5. **Regression Analysis Results for Predicting Item Difficulty for Officer Applicants**

	Raw regression weights					Item no. for difficulty estimate[a]	
Subtest	**Constant**	**Difficulty**	**Position**	R^2	**SEE**	**Middle**	**Last**
Verbal Analogies	.336976	.767917	-.007579	.83	.09	12	25
Arithmetic Reasoning	.257121	.967170	-.007078	.87	.07	12	25
Reading Comprehension	.408797	.785401	-.008505	.72	.06	12	25
Data Interpretation	.387264	.696786	-.015164	.81	.09	12	25
Word Knowledge	.385957	.642522	-.006076	.81	.07	12	25
Math Knowledge	.431834	.640306	-.004592	.72	.06	12	25
Mechanical Comprehension	.079527	1.010351	.008307	.69	.08	10	20
Electrical Maze	.276946	.589683	-.022614	.87	.07	10	20
Instrument Comprehension	.322572	.596583	-.019919	.72	.06	10	20
Block Counting	.213189	.819473	-.034267	.87	.08	10	20
Aviation Information	.151989	.999527	-.002567	.80	.06	10	20
	(-.087193)	(.884420)	(-.011303)	(.85)	(.05)	10	20
Rotated Blocks	.033552	1.021232	.003256	.96	.04	07	15
General Science	.112028	.934539	.001727	.85	.06	10	20
	(.050163)	(.750858)	(-.002430)	(.88)	(.05)	10	20
Hidden Figures	-.074246	1.165829	-.014815	.87	.08	07	15

Note. Values reported in parentheses for Aviation Information and General Science subtests are based on OTS cadet samples. Other values are for basic airmen samples.

[a] Difficulty estimates for the first position were computed for Item 1 in all subtests.

Item Record Viewing

Chart 1 demonstrates how the complete data for one item residing in the computer item bank would appear when accessed on a terminal. A block of data appears in such a way that the statistical information is lined up much as it is in the hard copy printout. Columns are lined up for ease in reading the data.

Additional Components of the Item Bank

The item data on tape, narrative text on diskette, and graphics in the card deck comprise the primary components of the item banking system. Secondary components are three documents intended to help test construction specialists use the item bank. These components are printed test booklets, hard copy of results of the classical item analysis and quintile analysis, and the taxonomy of item content. A variety of codes was designed to allow easy cross-referencing among the components. A list of the components with information on how they interrelate is provided below.

1. Data Tape. Each of the additional primary and secondary components is described with reference to how its item information relates to and is identifiable from variable codes on the data tape.

2. Item Text. For each AFOQT subtest, the text of narrative items in each booklet appears as a document file in a set of diskettes designed to be used on a microcomputer or uploaded to a mainframe. For each item the text is identified by the Content Area two-letter code (columns 1 - 2), Set Identifier (columns 3 - 4), and Item Number (columns 6 - 7).

3. Card Deck. All items that contain illustrations or special mathematical symbols are in the card deck. The Illustration Identifier (columns 34 - 37) refers to the 5" x 8" card on which the graphic appears. If an illustration has several items associated with it, these items will have the same illustration code. In the case where the illustration is the item, as in the EM, IC, RB, and HF subtests, an "X" appears in column 34 indicating an illustration with the same number as the Item Identification Code. In subtests AR, MK, and GS, the items that have special symbols appear on an illustration card. An "X" appears in column 34 for these items as well. Column 34 is blank for all other types of items.

4. Printed Booklets. The printed booklet contains the complete hard copy text of the test directions, items, illustrations, and layout. The booklet cover contains the Content Area, Set Identifier, and Booklet Identifier.

5. Item Analysis Printouts. The results of the classical item analysis and quintile analysis for each booklet were printed and assembled into separate documents for each subtest. A special purpose program was created so that the output was produced using a Laser printer on 8 1/2" x 11" paper, a document size more easily storable and accessible than the usual wide computer paper. The documents present all data in a concise, readable form, with the common items appearing first on one page, and the experimental items following on the next two pages. The format of the data permits quick scanning of the results of the analysis for an entire booklet.

6. Taxonomy of Content. The Content Category Identifier (columns 30 - 33) is a code representing the subject content of that item. Ten of the 16 subtests contained the kind of information capable of being categorized. The categorizing served the purpose of developing new items that were balanced to match the content in AFOQT Form O. Appendix A lists descriptions of the content categories and the codes that appear in columns 30 - 33.

Chart 1. **Complete Data For One Item In the Item Bank**[a]

Appearance of Data

```
AI01C12084092C CC20711850186 DP                 .93300  .81700  .15900
AI01C1210.5024-.3158  944-.2650 12460.6079 5055-.4288 1543-.1744 1477
AI01C122      -.1803 19  -.1616 24  0.4850104  -.2799 31  -.1119 29
AI01C123 .3519 .3402 .3272  822 45 2328 44 2937 43 3847 42 4859 33
AI01C124  9 20.0  4.3  5 11.3  2.4  3  6.9  1.4  0  0.0  0.0  2  6.0  0.9
AI01C125  7 15.5  3.3  8 18.1  3.8  4  9.3  1.9  4  9.5  1.9  1  3.0  0.4
AI01C126  8 17.7  3.8 13 29.5  6.2 27 62.7 13.0 28 66.5 13.5 28 84.8 13.5
AI01C127 13 28.8  6.2 10 22.7  4.8  4  9.3  1.9  3  7.1  1.4  1  3.0  0.4
AI01C128  8 17.7  3.8  8 18.1  3.8  5 11.6  2.4  7 16.6  3.3  1  3.0  0.4
AI01C129  0  0.0  0.0  0  0.0  0.0  0  0.0  0.0  0  0.0  0.0  0  0.0  0.0
```

Interpretation of Data

```
ID:  Content Area/ Set/ Subject Type/  Item Number/ Record Number
          AI        01  C (OTS-cadets)      12          0 to 9
```

```
RECORD 0:  Identification and IRT
AI01C12084092C CC20711850186 DP                 .93300  .81700  .15900

Booklet/ Key/ Speed/ Type/ Subject/ Size/ Test Date/ Form/ Cat./ a-b-c
 84092    C     no    C      OTS     207  11-85/1-86   no    DP    IRT
```

```
RECORDS 1 and 2:  Classical Item Analysis
  Response Option:  A          B    |     C     |     D          E
AI01C1210.5024-.3158  944-.2650 12460.6079 5055-.4288 1543-.1744 1477
AI01C122      -.1803 19  -.1616 24  0.4850104  -.2799 31  -.1119 29

Difficulty/   For Each Response:   Bis. Cor./ % Resp/ "T"
                                  Point Bis./  N   /  Blank
```

```
RECORD 3:  Difficulty Estimates and Quintile Sizes
AI01C123 .3519 .3402 .3272  822 45 2328 44 2937 43 3847 42 4859 33

Est. Diff: 1st, mid, last/  Per Quintile:  Low Score, High Score, N
```

```
RECORDS 4 - 9:  Quintile Statistics--Each Quintile
  Quintile:  1          2     |     3      |    4          5
AI01C124  9 20.0  4.3  5 11.3  2.4 | 3  6.9  1.4 | 0  0.0  0.0  2  6.0  0.9

Per Quintile:                        N
Record - Response            % That Quintile
  4  =  A                                   % Total
  5  =  B
  6  =  C
  7  =  D
  8  =  E
  9  =  Blank.
```

[a] Refer to Appendix C for File Layout Descriptions.

IV. DISCUSSION AND CONCLUSIONS

A test item bank like this one developed for the AFOQT achieves its purposes when all the information and parameters are ordered such that items may be efficiently cataloged, maintained, and modified, in particular when multiple forms and frequent testing are required. The efficiency and security of a computerized test item bank for the purpose of computer-assisted test construction (CATC) are well documented (Lee, Palmer, & Curran, 1988; Muiznieks & Dennis, 1979; Ree, 1978). Ree noted how time consuming and laborious it is to search large files of cards to select a few items and demonstrated that a CATC system not only permits a rapid search of many items, but can do so without clerical error and with maximum security. Other advantages include searching by identified parameters, saving of storage space, flexibility in terms of updating information and adding item categories, and increased capability (over cards) of displaying all item statistics.

The amount of data for an item that can be stored on a tape or disk is vast. The space availability encourages storage of more desirable information on items like those written for the AFOQT than would otherwise be recorded if the data had to be retrieved from enormous numbers of cards. Desirable data for each item include identifiers; item characteristics derived from both classical and IRT statistics; sample size and description; item text; keyed response; taxonomic classification; flags to other items which should not be used with the given item on the same subtest because of overlapping content; number of times and on what dates the item has been used operationally; and history of item modifications, if any (Lee, Palmer, & Curran, 1988). It would also be useful for item data to include information on the item construction phase, that is, whether the listed data were compiled from experimental or operational testing. Item characteristics such as "speeded" or "power," common or new, and whether the item has appeared in an operational booklet should be entered as well. Finally, printing information on typeface, pitch, spacing, and format of each item are necessary if future test forms are to be maximally parallel to the original form.

The new AFOQT item bank contains the essential components for automated test construction and contains most of the data suggested by the authors cited. Appendix C provides codes and explanations to facilitate retrieval of desired information. Further, the files can be updated to identify items selected for new forms, and additional data, if collected, can easily be inserted.

Item statistics are of particular concern to designers of test item banks. In an interesting report on item analysis presentation, Wainer (1988) recommends a CATC system that would capture and present item analysis results in more useful ways than did certain traditional outputs. He emphasized, for example, the substitution of graphs for quintile numeric data. The viewer would see at a glance whether or not the responses to the correct choice rise toward the higher quintiles, and the responses to the wrong choices descend or waver. The next step in providing a more sophisticated AFOQT item bank would be to employ the Macintosh SE hypercard or its equivalent (or future emulation on the IBM-DOS systems) as described by Wainer (1988). The test constructor would have a user-friendly screen and icons to bring up all pertinent data on the screen. This would allow, for example, a quintile graph to enhance the ease of checking for the desired increase in numbers who select the correct alternative as their total scores increase. Illustrations, text, and all statistics could also be viewed on the screen.

The increasing use of computerized item banking has generated some philosophical issues. Hsu and Sadock (1985), for example, expressed the concern that item classification must be meaningful and systematic, not dictated by the constraints of a quick retrieval system. They emphasized that assistance to users in selecting high-quality items should never be secondary to the advantages of efficient storage. This was also a concern of the AFOQT item bank project and the taxonomy was developed to be useful before consideration of how the item data were to be formatted for CATC.

The possibility of having to compromise when a great number of variables have to be considered is a danger of item banking, whether the data are computerized or not. For example, a test constructor may assemble items with a desired range of content, biserial correlations, and item difficulties, but the scoring key of the newly assembled test may not be balanced with respect to positions of the correct answers (i.e., A - E). If the positions of the alternatives are changed to balance the key, it may be difficult to justify generalizing the experimental results on the original forms. Similarly, equivalent distributions of item length and item format may have to be partially sacrificed in favor of appropriate item statistics.

Several limitations of computerized item banks have been observed. With respect to physical reproduction, some printing systems lack lower-case letters, have limited capability for reproducing pictorial items, and have difficulty with printing subscripts and superscripts (Muiznieks & Dennis, 1979). Two of these limitations did occur in the AFOQT item bank project; most figural items and math subtest items with special symbols, usually subscripts, were provided on cards. (However, references to these cards and all item data are included in the tapes and disks.)

The AFOQT item development project resulted in the storage of a large number of acceptable items for future new forms. Further, the item bank was developed in a manner which permits the number of items available for test construction to be increased easily. Many of the items that did not meet all the statistical requirements are potentially acceptable if appropriately revised, or, in some cases, administered to other samples. In the first instance, an item may have had a biserial correlation of zero instead of a negative correlation for one of the four wrong alternatives, but met all the other statistical requirements. In the second instance, the airmen samples, for example, may not have been the most appropriate group on which to field test some of the subtests. To insure that the contents of the item bank are fully utilized for test construction, it is recommended that opportunities be made for item revision and subtest re-administration.

The image of a computerized item bank as a "changing test folder" that accumulates information as items are revised or added is a compelling one. The AFOQT item bank developed for this project can exemplify this image. The number of items that met the discrimination criterion supports the conclusion that a sufficient number of acceptable items exist for the creation of new forms, the tape design permits changes to any test parameters, and the interface guides allow complete perusal of all item variables.

REFERENCES

Arth, T.O. (1985). Validation of the AFOQT for non-rated officers (AFHRL-TP-85-50, AD-A164 134). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Arth, T.O., & Skinner, M.J. (1986). Aptitude selectors for Air Force officer non-aircrew jobs. Proceedings of the 28th Annual Conference of the Military Testing Association (pp. 301-306). New London, CT.

Arth, T.O., Steuck, K.W., Sorrentino, C.T., & Burke, E.F. (in preparation). Air Force Officer Qualifying Test (AFOQT): Predictors of undergraduate pilot training and undergraduate navigator training success. Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Berger, F.R., Gupta, W.B., Berger, R.M., & Skinner, J. (in preparation). Air Force Officer Qualifying Test (AFOQT) Form P: Test manual. Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Berger, F.R., Gupta, W.B., Berger, R.M., & Skinner, J. (1988). Air Force Officer Qualifying Test (AFOQT) Form P: Test construction (AFHRL-TR-88-30, AD-A200 678). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Cowan, D.K., Barrett, L.E., & Wegner, T.G. (1989). Air Force Reserve Officer Training Corps selection system validation (AFHRL-TR-88-54). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Cowan, D.K., Barrett, L.E., & Wegner, T.G. (1990). Air Force Officer Training School selection system validation (AFRHL-TR-89-65). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Finegold, L.S., & Rogers, D.L. (1985). Relationship between Air Force Officer Qualifying Test scores and success in air weapons controller training (AFHRL-TR-85-13, AD-A158 162). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Gulliksen, H. (1950). Theory of mental tests (p. 15). New York: John Wiley and Sons, Inc.

Hsu, T.C., & Sadock, S.F. (1985). Computer-assisted test construction: The state of the art. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

Koplyay, J.B. (1981). Item analysis program (IAP) for achievement tests (AFHRL-TP-81-22, AD-A107 884). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Lee, W.M., Palmer, P., & Curran, L.T. (1988). Automated item banking and test development. (AFHRL-TP-88-40, AD-A205 870). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Mislevy, R.J., & Bock, R.D. (1984). BILOG II, Version 2.3: Item analysis and test scoring with binary logistic models. Morresville, IN: Scientific Software, Inc.

Muiznieks, B., & Dennis, J.R. (1979). A look at computer-assisted testing operations (No. 12e). Urbana, IL: Illinois University. The Illinois Series on Educational Application of Computers.

Ree, M.J. (1978). Automated test item banking (AFHRL-TR-78-13, AD-A054 626). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Rogers, D.L., Roach, B.W., & Wegner, T.G. (1986). Air Force Officer Qualifying Test Form O: Development and standardization (AFHRL-TR-86-24, AD-A172 037). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Skinner, M.J., & Ree, M.J. (1987). Air Force Officer Qualifying Test (AFOQT): Item and factor analysis of Form O (AFHRL-TR-86-68, AD-A184 975). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Wainer, H. (1988). The future of item analysis. (Research Report.) Princeton, NJ: Educational Testing Service.

Wesman, A.G. (1971). Writing the test item. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education.

APPENDIX A: TAXONOMY OF SUBTEST CONTENT

Subtest	Content code	Description
Verbal Analogies (12 booklet sets)		
	S	"Short" responses are single words
	L	"Long" responses are in the form: --- is to ---.
Arithmetic Reasoning (7 booklet sets)		
	G	Geometry: volume, area, circumference, triangles.
	P	Percentages
	D	Rates/distance, speeds, production rates
	A	Algebraic word problem. Many problems coded D, P, or R could be solved using an algebraic equation but are more likely solved by a simpler computation.
	R	Ratios, part-to-part and linear relationships.
Reading Comprehension (8 booklet sets)		
		Life Science
	LS A	Agriculture
	LSBO	Botany
	LS E	Ecology
	LS M	Medicine
	LS P	Physiology
	LS Z	Zoology
		Physical Science
	PS A	Astronomy
	PS C	Chemistry
	PSGP	Geography
	PSGL	Geology
	PSMA	Measurement
	PSMC	Mechanics
	PSMT	Meteorology
	PS O	Oceanography
	PS P	Physics

APPENDIX A: (Continued)

Subtest	Content code	Description
		Social Science
	SS A	Anthropology
	SSAR	Archeology
	SS B	Business
	SSEC	Economics
	SSED	Education
	SS H	History
	SSPH	Philosophy
	SSPO	Political Science
	SSPS	Psychology
	SS S	Sociology
		Art and Literature
	ALAR	Architecture
	AL A	Art
	AL G	Linguistics
	AL L	Literature
	AL M	Music
	AL S	Stories
Word Knowledge	(10 booklet sets)	
	A	Adjective/Adverb
	N	Noun
	V	Verb
Math Knowledge	(9 booklet sets)	
	E	"Equation" problem in elementary algebra
	F	"Factoring" problem in elementary algebra
	G	"Geometric" principles and rules
	P	"Properties" roots, proportions, functions, relationships
	AR	Arithmetic reasoning
Mechanical Comprehension		(16 booklet sets)
		Illustrated Items
	1	Springs
	2	Levers and leverage
	3	Transfer of rotational motion (gears)
	4	Rotational to linear motion

APPENDIX A: (Continued)

Subtest	Content code	Description
	5	Fluid and hydraulic systems
	6	Weights and pulleys
	7	Rotational range of motion
	8	Engines
	9	Tools
	M	Miscellaneous
		Non-illustrated Items
	1	Metal terminology
	2	Physics and physics terms
	3	Hardware
	4	Tools
	5	Gears and levers
	6	Fluid and fluid systems
	7	Cars
	8	Electrical comprehension
	M	Miscellaneous
Scale Reading	(7 booklet sets)	
		Each illustration is coded for four dimensions.
	Col 1	Scale
	E	Equal
	L	Log
	Col 2	Ruler Line
	S	Straight
	C	Curved
	Col 3	Number
	D	Decimal
	W	Whole
	Col 4	Scale Arrangement (lowest to highest scale value)
	E	East (scale reads from left to right)
	W	West (scale reads from right to left)
	N	North (scale reads from bottom to top)
	S	South (scale reads from top to bottom)

APPENDIX A: (Continued)

Subtest	Content code	Description
Instrument Comprehension		(7 booklet sets)
		Each illustration is coded for four dimensions.
	Col 1	Fuselage
	L	Level
	D	Dive
	C	Climb
	Col 2	Wing
	E	Even
	R	Right wing down
	L	Left wing down
	Col 3	Wing slant
	0	0°
	3	30°
	6	60°
	9	90°
	Col 4	Compass Direction
	1	North
	2	Northwest
	3	West
	4	Southwest
	5	South
	6	Southeast
	7	East
	8	Northeast

APPENDIX A: (Continued)

Subtest	Content code	Description
Aviation Information (7 booklet sets)		
		General
	GA	Aircraft
	GR	Regulations
	GO	Organizations
		Navigation/Communications
	NR	Radio
	NA	Airport
	M	Meteorology
		Aerodynamics
	AA	Aircraft
	AP	Performance
		Functions
	FO	Operations
	FP	Parts
		Definitions
	DT	Terms
	DP	Parts
		Hazards
	HA	Aircraft
	HC	Conditions
		Operations
	OF	Flying
	OL	Landing
	OC	Conditions
	ON	Navigation
	OP	Procedures

APPENDIX A: (Concluded)

Subtest	Content code	Description
General Science	(11 booklet sets)	
	AS	Astronomy
	BI	Biology
	CH	Chemistry
		Earth Science
	ES G	Geology/Geography
	ES M	Meteorology
	BP	Basic Physics
		Quantum Physics
	QP A	Atomics
	QP R	Radiation
		Electronics/Computers
	EC E	Electronics
	EC C	Computers
		Instruments/Measurement
	IM I	Instrumentation
	IM M	Measurement
	AI	Aviation Information

Note. This appendix provides the content category description and code for the 10 of the 16 subtests amenable to categorization.

APPENDIX B: INSTRUCTIONS FOR ITEM WRITERS

1. Follow the format of the sample items. All items are multiple choice with five alternatives (except for Instrument Comprehension, which has four alternatives).

2. Order item alternatives in ascending or descending order of length.

3. Order numeric alternatives in ascending or descending order.

4. Express items clearly in language appropriate for a high school reading level.

5. Avoid unnecessary repetition in the alternatives by including as much of the relevant information as possible in the item stem. For example, "The best way to estimate cost is through the use of:" is preferred to ending the stem with "is" and preceding all the alternatives with "through the use of."

6. Use the same terms and definitions consistently across items. For example, use either abbreviation or a whole word consistently.

7. Avoid absolutes such as "always" and "never."

8. All alternatives should have the same grammatical structure. For example, use the same tense throughout a question's alternatives; use one voice (active or passive) so that unusual structure doesn't give away key.

9. Avoid ambiguous or vague terms. For example, a specific time reference ("hourly," "monthly") is preferable to "frequently."

10. Avoid colloquialisms; use standard English.

11. Avoid inclusion of nonfunctional words or unnecessary detail to keep items as short and concise as possible.

12. Do not use "none of the above" as item alternatives.

13. If an item stem contains factual information, that information must be accurate.

APPENDIX B: (Concluded)

14. Write items such that there is only one correct answer possible among the alternatives. That is, provide a reasonable basis for response selection.

15. Insofar as possible, write items that reflect aspects of the work for which examinees are being tested. That is, the relevant reference is military jobs in the Air Force.

16. Avoid non-relevant clues to the correct response. Examples: making the correct alternative stand out by having it quite different from the other four in grammar, length, vocabulary, etc.; making it obvious by having the wrong alternatives appear to be silly and therefore transparently wrong.

17. Avoid sources of difficulty (e.g., unfamiliar language or symbols) that are not directly related to the content area tested.

18. Vary the difficulty of items. The number of items you are asked to write may not be enough to cover all levels of difficulty in the right proportions, and it is next to impossible to know the exact level of difficulty of an item prior to its testing. However, insofar as possible, write approximately one-third of your items to be low in difficulty, one-third to be of medium difficulty, and one-third to be of high difficulty.

19. The item stem should be informative to the point that the question is understood before reading the response alternatives.

20. No item will be accepted that contains controversial material regarding sensitive issues such as morality, religion, politics, ethnicity, or regionality.

APPENDIX C: FILE LAYOUT FOR AFOQT ITEM DATA TAPE

Record	Columns	Format	Variable description
0	1-2	A2	Content Area Identifier
	3-4	A2/I2	Booklet Set Number
	5	A1	Subject Type: A=Airmen, C=OTS cadets, R=ROTC
	6-7	A2/I2	Item Number
	8	A1/I1	Record Number "0"
	9-13	A5/I5	Booklet Number
	14	A1	Keyed Response: A-E or A-D
	15	A1	Speeded: F=Forward, B=Backward, Blank=Power
	16	A1	Item Type: C=Common, E=Experimental
	17	A1	Subject Type: A=Airmen, C=OTS, R=ROTC
	18-20	I3	Sample Size (for that testing)
	21-28	A8/I8	Testing Date: "MoYr" for first and last month of testing.
	29	A1	Operational Test Form: 1= P1, 2= P2, 3= Info Pac, 4= Common in P1+P2
	30-33	A4	Content Category Identifier[a]
	34-37	A4	Illustration Identifier "X"
	38-41	A4	Duplicate Item Identifier
	42-44	A3	Line Length (RC only)
	45		Blank
	46-72	3(F9.5)	IRT: a, b and c (overflow values = 9.99999)
	73		Blank
1	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "1"
	9-14	F6.4	Item Difficulty (Raw)
			Classical Item Analysis, Response A
	15-20	F6.4	Biserial Correlation (-1.0 set to -.9999)
	21-23	I3	Percent giving that response
	24-25	I2	T value that response
			Repeat 3 variables Responses B - E
	26-36		Response B
	37-47		Response C
	48-58		Response D
	59-69		Response E
	70-73		Blank
2	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "2"
	9-14	-	Blank
			Classical Item Analysis, Response A
	15-20	F6.4	Point Biserial Correlation
	21-23	I3	Number of Examinees giving that response
	24-25	-	Blank
			Repeat variables, Responses B - E
	26-36		Response B
	37-47		Response C
	48-58		Response D
	59-69		Response E

APPENDIX C: (Continued)

Record	Columns	Format	Variable description
3	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "3"
	9-25	3(F6.4)	Officer Item Difficulty Estimates First, Middle, and Last Position
	27-34		Test Score Range for Quintile 1
	27	-	Blank
	28-29	I2	Lowest Score that Quintile
	30-31	I2	Highest Score that Quintile
	32-34	I3	Number of Examinees that Quintile
	35-42		Test Score Range for Quintile 2
	43-50		Test Score Range for Quintile 3
	51-58		Test Score Range for Quintile 4
	59-66		Test Score Range for Quintile 5
	67-73		Blank
4	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "4"
			Quintile Statistics for Response A
	9-21		Statistics for Quintile 1
	9-11	I3	Number of Examinees
	12-16	F5.1	% Examinees that Quintile
	17-21	F5.1	% of Total Examinees
	22-34		Statistics for Quintile 2
	35-47		Statistics for Quintile 3
	48-60		Statistics for Quintile 4
	61-73		Statistics for Quintile 5
5	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "5"
			Quintile Statistics for Response B
	9-21		Statistics for Quintile 1
	9-11	I3	Number of Examinees
	12-16	F5.1	% Examinees that Quintile
	17-21	F5.1	% of Total Examinees
	22-34		Statistics for Quintile 2
	35-47		Statistics for Quintile 3
	48-60		Statistics for Quintile 4
	61-73		Statistics for Quintile 5
6	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "6"
			Quintile Statistics for Response C
	9-21		Statistics for Quintile 1
	9-11	I3	Number of Examinees
	12-16	F5.1	% Examinees that Quintile
	17-21	F5.1	% of Total Examinees
	22-34		Statistics for Quintile 2
	35-47		Statistics for Quintile 3
	48-60		Statistics for Quintile 4
	61-73		Statistics for Quintile 5

APPENDIX C: (Concluded)

Record	Columns	Format	Variable description
7	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "7"
			Quintile Statistics for Response D
	9-21		Statistics for Quintile 1
	9-11	I3	Number of Examinees
	12-16	F5.1	% Examinees that Quintile
	17-21	F5.1	% of Total Examinees
	22-34		Statistics for Quintile 2
	35-47		Statistics for Quintile 3
	48-60		Statistics for Quintile 4
	61-73		Statistics for Quintile 5
8	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "8"
			Quintile Statistics for Response E[b]
	9-21		Statistics for Quintile 1
	9-11	I3	Number of Examinees
	12-16	F5.1	% Examinees that Quintile
	17-21	F5.1	% of Total Examinees
	22-34		Statistics for Quintile 2
	35-47		Statistics for Quintile 3
	48-60		Statistics for Quintile 4
	61-73		Statistics for Quintile 5
9	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "9"
			Quintile Statistics for Blanks
	9-21		Statistics for Quintile 1
	9-11	I3	Number of Examinees
	12-16	F5.1	% Examinees that Quintile
	17-21	F5.1	% of Total Examinees
	22-34		Statistics for Quintile 2
	35-47		Statistics for Quintile 3
	48-60		Statistics for Quintile 4
	61-73		Statistics for Quintile 5

[a] See Appendix A for valid codes.

[b] When there is no E response, a "0" appears in the rows or columns alloted to the E response position.

APPENDIX D: AFOQT ITEM TAPE PROCEDURE: PROGRAM DESIGN AND DATA MANIPULATION

I. Create Programs (in Fortran)

1. Read original data tapes; count N per booklet; remove some variables to create the required shorter record length for Bilog; write all responses for each booklet to separate files.

2. Compute threshold means for inserting as link values for each content area for Bilog runs.

3. Conduct IAP (Item Analysis Program) to compute classical item statistics and quintile statistics.

4. Design "Collect Variables" program to read output from Item Analysis, Bilog and from data compiled from coding, such as content category codes.

II. Use Existing Programs for which Control Cards for Each Booklet Set Were Prepared

1. Write system Utility programs to sort files; copy them from one location to another (tapes, disk packs).

2. Run Bilog, written by Robert J. Mislevy and R. Darrell Bock, to compute IRT statistics. This program will not accept more than one set of data at a time, so the program was run separately on each booklet data set.

III. Manipulate Data Sets (Files)

1. Sixteen content areas with replications in SR and TR (reverse), and Cadets in GS, AI, and MC. Total replication sets = 21.

   Total administered booklet sets of the 16 AFOQT content areas = 176.

2. Bilog produces two data sets, one with the total printout, the second with the IRT data to be added to the final item bank.

IV. Test Programs and Verify Data

At each stage, every program was run and debugged many times. Frequent additions were made to improve the output in various ways. Output from the IAP was compared to a similar program developed at AFHRL (Koplyay, 1981). Since the final program to combine data used unique information for each content area, the information was carefully checked. Programs at each stage were run many times for checking output and making corrections. The numbers of programs and data sets listed in the next section are those retained from hundreds of data sets.

APPENDIX D: (Concluded)

V. Submit Programs and Create Files

Data management activities	Programs run	Files read	Files created	Files copied
1. Copy tapes onto mainframe	11			11
2. Sort files to verify that booklet responses are sequential	11	11	11	
3. Run Program 1 on each file	11	11	190	
4. Run Bilog on Set One of each content area set	23	23	46	
5. Run Program 2 on Bilog Set One output, printed output only	23	23	0	
6. Run Bilog on all other booklet sets	167	167	334	
7. Concatenate data from booklet sets				
Bilog Print data	on line	190	23	
Bilog IRT	interactive	190	23	
Booklet Response Data		190	23	
8. Delete all small files	8			
9. Run Program 3, IAP				
Print file format for paper output	23	23	23	
File format for data tape		23	23	
10. Sort File for data tape so all cards for each item are in sequence	23	23	23	
11. Run Program 4, combine all data	23	23	23	
12. Copy files onto unlabeled tape	1			46
13. Back up all files on labeled tape	1			161
14. Delete remaining files	1			
Totals	326	897	742	218

<u>APPENDIX E</u>: NUMBER OF EXPERIMENTAL ITEMS MEETING ITEM DISCRIMINATION INDEX ACCEPTABILITY CRITERIA BY SUBTEST AND BY SAMPLE

	Booklet number										
	1	2	3	4	5	6	7				
Subtest	8	9	10	11	12	13	14	15	16	Subtotal	Total
Verbal Analogies											
Airmen -- Phase 1	18	25	16	11	12	12	8			102	
Airmen -- Phase 2	16	30	21	17	13					97	199
Arithmetic Reasoning											
Airmen -- Phase 1	28	24	24	26	27	27	27			183	183
Reading Comprehension											
Airmen -- Phase 1	25	17	15	24	23	20	20			144	
Airmen -- Phase 2	30									30	174
Data Interpretation											
Airmen -- Phase 1	25	20	32	25	26	31	29			188	
Airmen -- Phase 2	13									13	201
Word Knowledge											
Airmen -- Phase 1	25	22	15	22	14	19	16			133	
Airmen -- Phase 2	28	20	21							69	202
Math Knowledge											
Airmen -- Phase 1	18	26	22	25	16	19	22			148	
Airmen -- Phase 2	13	15								28	176
Mechanical Comprehension											
Airmen -- Phase 1	14	10	11	5	9	16	11			76	
Airmen -- Phase 2	15	12	12	9	14	9	11	8	7	97	97
OTS -- Phase 2	22	18	22	15						77	
ROTC -- Phase 2					18	28	16	15	12	89	
Cadets -- Total											166
Electrical Maze											
Airmen -- Phase 1	37	33	37	30	37	37	37			248	248

APPENDIX E: (Concluded)

	Booklet number										
	1	2	3	4	5	6	7				
Subtest	8	9	10	11	12	13	14	15	16	Subtotal	Total
Scale Reading											
Airmen -- Phase 1	25	26	20	25	29	28	26			179	179
Instrument Comprehension											
Airmen -- Phase 1	33	37	36	38	38	35	36			253	253
Block Counting											
Airmen -- Phase 1	13	19	27	13	31	24	26			153	
Airmen -- Phase 2	35	25								60	213
Table Reading											
Airmen -- Phase 1	21	24	28	26	23	26	37			185	185
Aviation Information											
Airmen -- Phase 1	6	7	9	7	9	8	6			52	52
OTS -- Phase 1	24	22	28	26	26	27	24			177	177
Rotated Blocks											
Airmen -- Phase 1	30	25	19	27	26	25	28			180	180
General Science											
Airmen -- Phase 1	14	8	10	10	4	3	6			55	
Airmen -- Phase 2	7	10	9	8						34	89
OTS -- Phase 1	13	15	15	14	12	9	7			85	
ROTC -- Phase 2	7	14	9	11						41	
Cadets -- Total											126
Hidden Figures											
Airmen -- Phase 1	39	35	40	38	41	41	40			274	274